Deep Neural Network Bottleneck Features For Generalized Variable Parameter HMMs

Xurong Xie[1,3]*, Rongfeng Su*[1,3]*, Xunying Liu*[1,2] *& Lan Wang*[1,3]

[1]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[2]Cambridge University Engineering Dept, Trumpington St., Cambridge, CB2 1PZ U.K.
[3]The Chinese University of Hong Kong, Hong Kong, China

xr.xie@siat.ac.cn, rf.su@siat.ac.cn, xl207@cam.ac.uk, lan.wang@siat.ac.cn

Abstract

Recently deep neural networks (DNNs) have become increasingly popular for acoustic modelling in automatic speech recognition (ASR) systems. As the bottleneck features they produce are inherently discriminative and contain rich hidden factors that influence the surface acoustic realization, the standard approach is to augment the conventional acoustic features with the bottleneck features in a tandem framework. In this paper, an alternative approach to incorporate bottleneck features is investigated. The complex relationship between acoustic features and DNN bottleneck features is modelled using generalized variable parameter HMMs (GVP-HMMs). The optimal GVP-HMM structural configuration and model parameters are automatically learnt. Significant error rate reductions of 48% and 8% relative were obtained over the baseline multi-style HMM and tandem HMM systems respectively on Aurora 2.

Index Terms: generalized variable parameter HMM, deep neural network, bottleneck features, robust speech recognition

1. Introduction

Recently deep neural networks (DNNs) have become increasingly popular for acoustic modelling in automatic speech recognition (ASR) systems [1, 2, 3, 4, 5, 6, 7, 8]. In order to incorporate DNNs, or multi-layer perceptrons (MLPs) in general, into HMM based acoustic models, two approaches can be used. The first uses a hybrid architecture that estimates the HMM state emission probabilities using DNNs [9]. The second approach uses an MLP or DNN as a feature extractor, trained to produce phoneme posterior probabilities. The resulting probabilistic features [10], or bottleneck features [11] are used to train standard GMM-HMMs in a tandem fashion. As these features capture additional information complementary to standard front-ends, they are often combined in tandem systems.

One important issue associated with the tandem HMM approach is the appropriate method used to combine the conventional and bottleneck features. The precise nature of the relationship between the two is highly complex. Compared with the standard front-ends, bottleneck features provide a different view of the same speech signals. Certain correlation can therefore exist between the two. At the same time, complementary information characterizing the underlying hidden factors influencing the surface acoustic realization are also implicitly learnt by bottleneck features. They are propagated into HMMs as additional cues and constraints to improve discrimination. The standard approach augments the conventional front-ends with bottleneck features in a concatenated form. More advanced approaches that explicitly approximate the correlation between them using linear, affine transformations have also been proposed [12, 13].

In order to better capture the complex relationship between standard acoustic and bottleneck features, techniques motivated by speech production that can fully exploit the hidden variability in the bottleneck features may be used. Along this line, an alternative method to incorporate bottleneck features into a tandem system is proposed in this paper. DNN bottleneck features are used as influence factors to directly introduce controllability to the underlying generative acoustic models that are based on generalized variable parameter HMMs (GVP-HMMs) [14, 15, 16, 17, 18]. The continuous trajectories of optimal HMM parameters against the time-varying hidden factors in the bottleneck features are modelled using polynomial functions. Their effects on the acoustic parameters are automatically learnt by locally optimized polynomial parameters and degrees. Using the proposed GVP-HMM tandem approach, significant error rate reductions of 48% and 8% relative were obtained over the multi-style baseline HMM and tandem HMM systems respectively on Aurora 2.

The rest of this paper is organized as follows. Generalized variable parameter HMMs and an associated efficient complexity control technique are introduced in section 2. Deep neural networks and bottleneck features are reviewed in section 3. A range of GVP-HMM systems using various modelling configurations are described in section 4. In section 5 various GVP-HMM systems using DNN bottleneck features are evaluated on Aurora 2. Section 6 is the conclusion and future research.

2. Generalized Variable Parameter HMMs

Generalized variable parameter HMMs (GVP-HMMs) [14, 15, 16, 17] explicitly model the parameter trajectories of optimal Gaussian components, or more compact tied linear transformations, that vary with respect to some influence factors. In this paper, trajectories of Gaussian means and variances are used.

2.1. Model Definition

For a D dimensional observation $\boldsymbol{o}_t$ emitted from Gaussian mixture component m, assuming Pth order polynomials modelling a total of N regression variables are used, the form of

This work is supported by National Natural Science Foundation of China (NSFC 61135003), National Fundamental Research Grant of Science and Technology (973 Project: 2013CB329305) Shenzhen Fundamental Research Program JC01005280621A, JCYJ20130401170306806.

GVP-HMMs considered in this paper is given by

$$\boldsymbol{o}^{(t)} \sim p\left(\boldsymbol{o}^{(t)}; \boldsymbol{\mu}^{(m)}(\mathbf{v}_t), \boldsymbol{\Sigma}^{(m)}(\mathbf{v}_t)\right). \quad (1)$$

$\mathbf{v}_t^\top$ is a $(P \times N + 1)$ dimensional Vandermonde vector [19],

$$\mathbf{v}_t^\top = \left[1,\ \tilde{\mathbf{f}}_{t,1},\ ..., \tilde{\mathbf{f}}_{t,p},\ ..., \tilde{\mathbf{f}}_{t,P}\right]^\top. \quad (2)$$

and its N dimensional pth order subvector is defined as $\tilde{\mathbf{f}}_{t,p} = [v_{t,1}^p,\ ..., v_{t,j}^p,\ ..., v_{t,N}^p]^\top$, where $v_{t,j}$ is the jth element of an N dimensional factor vector Gaussian parameters are conditioned on at frame t, for example, the DNN bottleneck features,

$$\vec{\mathbf{f}}_t = [v_{t,1},\ ...,\ v_{t,j}, ...,\ v_{t,N}]^\top. \quad (3)$$

$\boldsymbol{\mu}^{(m)}(\cdot)$ and $\boldsymbol{\Sigma}^{(m)}(\cdot)$ are the P^{th} order mean and covariance trajectory polynomials of component m respectively. When diagonal covariances are used, the trajectories of the i^{th} dimension of the mean and variance parameters are computed as

$$\begin{aligned} \mu_i^{(m)}(\mathbf{v}_t) &= \mathbf{v}_t \cdot \mathbf{c}^{(\mu_i^{(m)})} \\ \sigma_{i,i}^{(m)}(\mathbf{v}_t) &= \breve{\sigma}_{i,i}^{(m)} \mathbf{v}_t \cdot \mathbf{c}^{(\sigma_{i,i}^{(m)})} \end{aligned} \quad (4)$$

where $\mathbf{c}^{(\cdot)}$ is a $(P \times N + 1)$ dimensional polynomial coefficient vector and $\breve{\sigma}_{i,i}^{(m)}$ is the conventional HMM variance estimate.

As a natural form of generative model inspired by speech production, a range of factors influencing the acoustic realization of speech have been investigated in previous research using GVP-HMMs, or their precursors based on more restricted forms of parameter trajectories, such as multiple regression HMMs (MR-HMM) [20] and variable parameter HMMs (VP-HMM) [21, 22]. These acoustic factors include prosodic features [20], environment noise condition represented by the signal-to-noise ratio (SNR) [14, 15, 16, 17, 18, 21, 22], and more recently articulatory features for speech synthesis [23].

GVP-HMMs share the same instantaneous adaptation power and good controllability as MR-HMMs and VP-HMMs. For any variability indicated by the factor vector, e.g. the bottleneck features, or SNR level, present or unseen in the training data, GVP-HMMs can instantly produce the matching HMM model parameters by-design without requiring any multi-pass decoding and adaptation process.

2.2. Parameter Estimation for GVP-HMMs

For the form of GVP-HMMs of equation (1) the associated ML auxiliary function is given by [14, 15, 24],

$$\mathcal{Q}_{\mathsf{ml}}^{\mathsf{GVP}}(\theta, \tilde{\theta}) = \sum_{m,t} \gamma_m(t) \log p\left(\boldsymbol{o}^{(t)}; \boldsymbol{\mu}^{(m)}(\mathbf{v}_t), \boldsymbol{\Sigma}^{(m)}(\mathbf{v}_t)\right) \quad (5)$$

where $\gamma_m(t)$ is the posterior probability of frame $\boldsymbol{o}_t$ being emitted from component m at a time instance t.

Combining the above with equations (1) and (4), the corresponding parts of the above auxiliary function associated with the polynomial coefficient vectors of the Gaussian mean and variance trajectories respectively can be re-arranged into convex quadratic forms,

$$\begin{aligned} \mathcal{Q}_{\mathsf{ml}}^{(\mu_i^{(m)})}(\theta, \tilde{\theta}) &= -\frac{1}{2}\mathbf{c}^{(\mu_i^{(m)})\top}\mathbf{U}^{(\mu_i^{(m)})}\mathbf{c}^{(\mu_i^{(m)})} \\ &\quad + \mathbf{k}^{(\mu_i^{(m)})}\mathbf{c}^{(\mu_i^{(m)})} + \text{const} \\ \mathcal{Q}_{\mathsf{ml}}^{(\sigma_{i,i}^{(m)})}(\theta, \tilde{\theta}) &= -\frac{1}{2}\mathbf{c}^{(\sigma_i^{(m)})\top}\mathbf{U}^{(\sigma_i^{(m)})}\mathbf{c}^{(\sigma_i^{(m)})} \\ &\quad + \mathbf{k}^{(\sigma_i^{(m)})}\mathbf{c}^{(\sigma_i^{(m)})} + \text{const}' \end{aligned} \quad (6)$$

where the constant terms independent of the coefficient vectors $\mathbf{c}^{(\cdot)}$ can be ignored. Setting the above gradients against the respective polynomial coefficient vectors to zero, the following ML solutions of the coefficient vectors can then be derived

$$\begin{aligned} \hat{\mathbf{c}}^{(\mu_i^{(m)})} &= \mathbf{U}^{(\mu_i^{(m)})-1}\mathbf{k}^{(\mu_i^{(m)})} \\ \hat{\mathbf{c}}^{(\sigma_{i,i}^{(m)})} &= \mathbf{U}^{(\sigma_{i,i}^{(m)})-1}\mathbf{k}^{(\sigma_{i,i}^{(m)})} \end{aligned} \quad (7)$$

and the sufficient statistics are

$$\begin{aligned} \mathbf{U}^{(\mu_i^{(m)})} &= \sum_t \gamma_m(t)\sigma_{i,i}^{(m)-1}(\mathbf{v}_t)\mathbf{v}_t^\top\mathbf{v}_t \\ \mathbf{k}^{(\mu_i^{(m)})} &= \sum_t \gamma_m(t)\sigma_{i,i}^{(m)-1}(\mathbf{v}_t)o_i^{(t)}\mathbf{v}_t^\top \\ \mathbf{U}^{(\sigma_{i,i}^{(m)})} &= \sum_t \gamma_m(t)\breve{\sigma}_{i,i}^{(m)}\mathbf{v}_t^\top\mathbf{v}_t \\ \mathbf{k}^{(\sigma_{i,i}^{(m)})} &= \sum_t \gamma_m(t)\left(o_i^{(t)} - \mu_i^{(m)}(\mathbf{v}_t)\right)^2\mathbf{v}_t^\top \end{aligned} \quad (8)$$

2.3. Model Complexity Control for GVP-HMMs

An important issue associated with GVP-HMMs is the appropriate polynomial degree to use. The use of higher degree polynomials can result in severe over-fitting and oscillation [25]. In addition, the precise form of individual parameter trajectories should be in line with the nature of the distinct effects imposed on them by the influencing factors. In order to more flexibly capture these complex, potentially locally varying effects and improve robustness, the optimal polynomial degrees of Gaussian mean and variance trajectories can be automatically determined at local level using complexity control techniques [18].

In Bayesian learning, when no prior knowledge over model structures $\{\mathcal{M}\}$ is available, the optimal model structure or complexity, is determined by maximizing the *evidence*,

$$p(\mathcal{O}|\mathcal{W}, \mathcal{M}) = \int p(\mathcal{O}|\theta, \mathcal{W}, \mathcal{M})p(\theta|\mathcal{M})\mathrm{d}\theta \quad (9)$$

where θ is a parameterization of $\mathcal{M}$, $\mathcal{O} = \{\boldsymbol{o}_1, ..., \boldsymbol{o}_\mathcal{T}\}$ is a training data set of $\mathcal{T}$ frames and $\mathcal{W}$ the reference transcription.

For standard HMMs and GVP-HMMs, it is computationally intractable to directly compute the evidence in equation (9). To handle this problem, an efficient approximation using the BIC style first order asymptotic expansion [26] of a lower lower bound [18, 27, 28, 29] of the evidence integral can be used. The optimal model complexity is determined by

$$\hat{\mathcal{M}} = \arg\max_{\mathcal{M}}\left\{\mathcal{Q}_{\mathsf{ml}}^{(\mathcal{M})}(\hat{\theta}, \tilde{\theta}) - \rho \cdot \frac{k}{2}\log\mathcal{T}\right\}. \quad (10)$$

where the ML auxiliary functions associated with Gaussian mean and variance trajectory parameters given in equation (6) evaluated at the optimal model parameters $\hat{\theta}$ using the statistics given in equations (7) and (8). k denotes the number of free parameters in $\mathcal{M}$ and ρ is a tunable penalty term [30].

When determining the optimal order for a particular polynomial associated with the ith dimension of the mth Gaussian component in the system, $\mu_i^{(m)}(\cdot)$, for example, the above statistics in equation (8) are accumulated for the highest order P_{max} being considered. The corresponding statistics for any other order $0 \leq P^{(\mu_i^{(m)})} < P_{\mathsf{max}}$ can be derived by taking the associated submatrices or subvectors from the full matrix statistics accumulated for P_{max}. Using these statistics and the ML

solutions in equation (7), the ML auxiliary function associated with $\mu_i^{(m)}(\cdot)$ in equation (6), can be efficiently evaluated at the optimum for each candidate polynomial degree. The number of free parameters (polynomial coefficients) in the BIC metric of equation (10) is $k = P^{(\mu_i^{(m)})}+1$. The number of frame samples for the current Gaussian is computed as the component level occupancy counts $\mathcal{T}^{(m)} = \sum_{t,m} \gamma_m(t)$. The same approach can also be used to determine the optimal degree of Gaussian variance polynomials by evaluating the respective auxiliary functions with their respective sufficient statistics to compute the metric in equation (10).

3. DNN Bottleneck Features

Bottleneck features are normally generated from a narrow hidden layer of an MLP that is trained to predict phonemes or phoneme states. Compared with the size of other layers, this hidden layer has a significantly smaller number of hidden units [11]. This narrow layer introduces a constriction in the network while retaining the information useful to classification in the resulting low dimensional features extracted via a non-linear and discriminative transformation.

In this paper the bottleneck features used for tandem HMM systems are extracted from deep neural network (DNN) multilayer perceptrons (MLP) [1, 2, 3, 4]. DNNs are MLPs with many hidden layers. The inputs are formed from a stacked set of adjacent frames of the acoustic feature for each time instance. Within each hidden layer, the input to each unit is computed as a linearly weighted sum of the outputs from the previous layer. Each hidden node transforms its input with a sigmoid activation to achieve non-linearity. An softmax output activation function is used at the output layer to compute the posterior probability of phonemes or phoneme state targets. In all the experiments of this paper, a pretrained DNN consisting of six hidden layers is used. The first five layers have a total of 512 hidden nodes while the last bottleneck layer has 26 units. The network is trained on inputs formed by splicing 11 frames of 39 dimensional MFCC features together. The layer-by-layer RBM based pre-training implemented in the Kaldi toolkit [31] was used.

Following DNN training 26 dimensional bottleneck features are extracted and decorrelated using PCA. For the baseline tandem HMM systems, they are appended to standard MFCC features to form the tandem feature vector. Prior to recognition, tandem GMM-HMMs are then trained based on the new concatenated tandem features. For GVP-HMM systems, these are used as the input factor vectors at each frame to estimate continuous trajectories of Gaussian mean and variance parameters. An extended version of the HTK toolkit [32] was used to train various GVP-HMM systems.

4. Using DNN Bottleneck Features In GVP-HMMs and Tandem GVP-HMMs

In order to adjust the trade-off between modelling resolution, robustness and computational efficiency, a range of GVP-HMM configurations may be considered to incorporate DNN bottleneck features. Description of these GVP-HMM variant systems' configurations and the number of parameters used for the standard Aurora 2 task are shown in table 1. 39 dimensional standard MFCC features including the first and second order differentials were used. All the baseline GVP-HMMs with no complexity control used 2nd degree polynomials for all parameter trajectories, as suggested in [21, 22]. The penalty term in the complexity control metric of equation (10) was fixed as $\rho = 1$ in all experiments. For all parameter polynomials the range of candidate degree to consider is $[0, 5]$.

Baseline HMM and tandem HMM systems: In the first 3 lines of table 1, the number of parameters for the multi-style [33] trained baseline HMM system and two tandem HMM systems are shown. The second tandem HMM system, "tandem†", used 18 Gaussians per state thus has a model complexity comparable to the other complexity controlled GVP-HMM systems in the table. The Gaussian parameters of these baseline HMM or tandem HMM systems were trained on standard MFCC or tandem features while no parameter trajectory modelling was used.

Model Type	System	Parm Poly mean	Parm Poly var	Com Ctrl	#Parm
HMM	mcond	-	-	-	79K
	tandem				132K
	tandem†				396k
GVP-HMM	mean	√	×	×	2.15M
				√	272K
	mv	√	√	×	2.27M
				√	362K
tandem GVP-HMM	mean	√	×	×	2.2M
				√	298K
	mv	√	√	×	2.32M
				√	406K

Table 1: Description of the baseline multi-style HMM, tandem HMM systems, GVP-HMM and tandem GVP-HMM systems on Aurora 2 in terms of model configurations and the number of parameters. Following the setting of previous works [21, 17, 18], all systems used 6 Gaussians per state except the "tandem†" baseline system used 18 Gaussians per state.

GVP-HMM systems: In the second section of table 1, a total of four GVP-HMM modelling configurations, denoted as "mean" and "mv" respectively, which use trajectory modelling for Gaussian component means using the DNN bottleneck features as the factor input in equations from (1) to (3), with the further options of using variance trajectories conditioned on the SNR variable, and with or without applying the model selection technique presented in section 2.3, are shown from the 4th to 7th line in table 1. As expected, using the standard GVP-HMMs with no complexity control on the 26 dimensional bottleneck features results in a massive increase in model parameters. Determining the optimal degrees for parameter trajectory polynomials using the model selection method of section 2.3 significantly reduced the model complexity by over to 80%.

Tandem GVP-HMM systems: In the last section of table 1, four comparable tandem variants of the above four GVP-HMM systems are shown. In these tandem GVP-HMM systems, the DNN bottleneck features are not only used as the input factor vectors to estimate the continuous trajectories of Gaussian parameters in the acoustic feature subspace, but also used as normal features to train the standard mean and variance parameters in the bottleneck feature subspace. For example, the final mean vector of component m at time instance t is thus computed as

$$\vec{\boldsymbol{\mu}}_t^{(m)} = [\boldsymbol{\mu}_{\mathsf{GVP}}^{(m)}(\mathbf{v}_t^{\mathsf{BN}}),\ \boldsymbol{\mu}_{\mathsf{BN}}^{(m)}]^\top. \tag{11}$$

where the $\boldsymbol{\mu}_{\mathsf{GVP}}^{(m)}(\mathbf{v}_t^{\mathsf{BN}})$ is the mean subvector trajectory taking a Vandermonde vector input $\mathbf{v}_t^{\mathsf{BN}}$ constructed using the 26 dimensional DNN bottleneck features, as described in section 2.1.

$\boldsymbol{\mu}_{\mathsf{BN}}^{(m)}$ is the remaining static mean subvector estimated using the bottleneck features. These tandem GVP-HMMs are expected to draw strength from both the conventional tandem and GVP-HMM based approaches to fully exploit the complementary information in the DNN bottleneck features.

5. Experiments and Results

In this section, the performance of various GVP-HMM systems using DNN bottleneck features are evaluated the Aurora 2 task. The Aurora 2 database contains different noisy conditions. During the experiments, 420 utterances from each of four different SNR conditions (-5dB, 5dB, 15dB, 25dB) of noise environments of subway, babble, car and exhibition were used to train all the systems, while 1000 utterances selected from each noise environment at 0dB, 5dB, 10dB, 15dB and 20dB SNR respectively were used for word error rate (WER) evaluation.

Noise Type	System	Com Ctrl	0dB	5dB	10dB	15dB	20dB	Ave
subway	mcond	-	21.25	7.55	3.78	2.36	2.27	7.44
	mean	×	15.08	5.59	3.01	2.12	1.23	5.41
		√	14.28	4.61	2.15	1.47	1.26	4.75
	mv	×	20.94	8.78	4.91	3.72	3.62	8.39
		√	12.59	4.51	2.00	1.44	1.01	**4.31**
babble	mcond	-	30.47	12.09	6.53	4.59	4.08	11.55
	mean	×	32.16	10.10	3.99	2.33	1.72	10.06
		√	26.96	8.04	2.63	1.90	1.15	8.14
	mv	×	38.42	14.33	6.38	4.35	3.51	13.40
		√	23.70	7.38	2.57	1.66	1.15	**7.29**
car	mcond	-	22.88	9.42	4.29	3.58	2.78	8.95
	mean	×	17.72	9.70	4.08	2.49	1.99	7.20
		√	15.63	6.55	2.82	2.10	1.90	5.80
	mv	×	26.01	14.36	6.03	4.10	3.26	10.75
		√	13.90	6.37	2.85	2.22	1.87	**5.44**
exhibition	mcond	-	23.46	10.27	4.91	3.09	2.72	8.89
	mean	×	14.60	5.86	2.99	1.54	1.23	**5.24**
		√	15.22	6.29	3.05	2.13	1.88	5.71
	mv	×	16.20	6.60	3.15	1.54	1.05	5.71
		√	15.68	6.26	2.78	1.45	0.96	5.43

Table 2: WER performance of GVP-HMM systems using DNN bottleneck features on Aurora 2 test set A of four noise types. All systems used the same naming conventions as in table 1.

The WER performance of the multi-style HMM baseline, "mcond", and various GVP-HMM systems shown from the 4th to 7th line of table 1 are given in table 2. The following trends can be found in the table. First, the use of DNN bottleneck features gave significant WER reductions for all GVP-HMM modelling configurations across various noise types over the "moncon" HMM baseline. Second, as expected, using the model selection technique of section 2.3, in addition to the model size compression shown previously in table 1, an average WER reduction of 2.41% absolute (29% relative) was obtained over various standard GVP-HMM systems with no complexity control. Third, combined with model complexity control, the use of variance trajectory polynomials gave further improvements over using mean trajectory modelling only. Using the best GVP-HMM systems highlighted in bold in table 2, an average WER reduction of 3.64% absolute (40% relative) over the multi-style MFCC feature trained baseline "mcond" HMM system was obtained. However, all of these four GVP-HMM systems were outperformed by the baseline tandem HMM system shown in the 1st line of each noise specific section in table 3.

The WER performance of the two baseline multi-style trained tandem HMM systems, "tandem" and "tandem†", and various tandem GVP-HMM systems shown from the 8th to 11th line in the bottom section of table 1 are given in table 3. Consistent with the trends found in table 2, every complexity controlled tandem GVP-HMM system in table 3 outperformed its comparable baseline using no complexity control. The use of variance trajectory modelling also gave further small reductions in WER. Using the best complexity controlled tandem GVP-HMM "mv" system highlighted in bold in table 3, an average WER reduction of 4.38% absolute (48% relative), and 0.4% absolute (8% relative) over the multi-style baseline "mcond" system of table 2, and the baseline "tandem" HMM system of table 3 respectively were obtained. Similar consistent improvements were also obtained over the more complex baseline "tandem†" system with a comparable number of parameters as shown in table 1, and a third baseline tandem HMM system using the bottleneck features extracted from a DNN trained on concatenated MFCC and SNR features.

Noise Type	System	Com Ctrl	0dB	5dB	10dB	15dB	20dB	Ave
subway	tandem	-	11.85	4.24	2.33	1.50	1.01	4.19
	tandem†	-	10.62	3.93	2.27	1.41	0.98	3.84
	mean	×	11.05	3.96	2.39	1.60	1.47	4.09
		√	11.08	3.99	2.15	1.41	0.86	3.90
	mv	×	10.99	4.05	2.70	2.18	2.24	4.43
		√	10.22	3.87	2.03	1.57	0.89	**3.72**
babble	tandem	-	20.77	7.10	3.14	1.69	1.12	6.76
	tandem†	-	20.65	6.77	2.78	1.51	1.03	6.55
	mean	×	21.13	7.50	3.30	1.66	1.33	6.98
		√	20.31	6.83	2.78	1.51	1.03	6.49
	mv	×	20.86	7.62	3.54	2.09	1.93	7.21
		√	20.19	6.80	2.81	1.45	1.12	**6.47**
car	tandem	-	10.86	4.81	3.12	2.16	1.60	4.51
	tandem†	-	11.40	4.66	3.09	2.13	1.54	4.56
	mean	×	10.26	4.93	2.73	2.31	1.81	4.41
		√	10.38	4.69	2.79	1.98	1.75	4.32
	mv	×	11.22	5.11	3.09	2.34	1.93	4.74
		√	9.99	4.63	2.82	1.97	1.66	**4.21**
exhibition	tandem	-	14.29	6.51	3.12	1.85	1.45	5.44
	tandem†	-	14.35	6.05	2.84	1.94	1.45	5.33
	mean	×	13.67	5.95	2.84	1.76	1.36	5.12
		√	13.61	6.20	2.72	1.70	1.27	5.10
	mv	×	15.03	5.55	3.05	2.19	1.91	5.55
		√	12.96	5.99	2.53	1.73	1.27	**4.90**

Table 3: WER performance of tandem GVP-HMM systems using DNN bottleneck features on Aurora 2 test set A. All systems used the same naming conventions as in table 1.

6. Conclusion

An alternative approach to incorporate bottleneck features into a tandem system using generalized variable parameter HMMs is investigated in this paper. The complementary information characterizing the hidden factors influencing the surface acoustic realization implicitly learnt by bottleneck features are exploited to improve controllability and robustness. The proposed technique significantly reduced the error rate by 48% and 8% relative over the baseline multi-style HMM and tandem HMM systems respectively on Aurora 2. Future research will focus on using bottleneck features to model the trajectories of more efficient feature space transforms [17].

7. References

[1] F. Side, G. Li, and D. Yu (2011). "Conversational speech transcription using context-dependent deep neural networks", in *Proc. ISCA INTERSPEECH2011*, pp. 437-440, Florence, Italy.

[2] D. Yu and M. L. Seltzer (2011). "Improved Bottleneck Features Using Pretrained Deep Neural Networks", in *Proc. ISCA INTERSPEECH2011*, pp. 237-240, Florence, Italy.

[3] G. Dahl, D. Yu, L. Deng, and A. Acero (2012). "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition", in *Proc. IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30C42, jan 2012.

[4] G. E. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kinsbury (2012). "Deep neural networks for acoustic modeling in speech recognition", *IEEE Signal Processing Magazine*, pp. 2-17, nov 2012.

[5] M. Seltzer, D. Yu and Y. Wang (2013). "An Investigation Of Deep Neural Networks For Noise Robust Speech Recognition", in *Proc. IEEE ICASSP2013*, pp. 7398-7402, Vancouver, BC, Canada.

[6] S. Thomas, M. L. Seltzer, K. Church, and H. Hermansky (2013). "Deep neural network features and semi-supervised training for low resource speech recognition", in *Proc. IEEE ICASSP2013*, pp. 6704-6708, Vancouver, BC, Canada.

[7] P. Bell, P. Swietojanski and S. Renals (2013). "Multi-level adaptive networks in tandem and hybrid ASR systems", in *Proc. IEEE ICASSP2013*, pp. 6975-6979, Vancouver, BC, Canada.

[8] K. M. Knill, M. J. F. Gales, S. P. Rath, P. C. Woodland, C. Zhang and S.-X. Zhang (2013). "Investigation of multilingual deep neural networks for spoken term detection", in *Proc. IEEE ASRU2013*, pp. 138-143, Olomouc, Czech Republic.

[9] H. A. Bourlard and N. Morgan (1993). "Connectionist Speech Recognition: A Hybrid Approach", Kluwer Academic Publishers, Norwell, MA, USA, 1993.

[10] H. Hermansky, D. Ellis and S. Sharma (2000). "Tandem connectionist feature extraction for conventional HMM systems", in *Proc. IEEE ICASSP2000*, vol. 3, pp. 1635-1638, Istanbul, Turkey.

[11] F. Grezl, M. Karafiat, S. Kontar and J. Cernocky (2007). "Probabilistic and bottle-neck features for LVCSR of meetings", in *Proc. IEEE ICASSP2007*, Vol. 4, pp. 757-760, Honolulu, Hawaii, USA.

[12] J. Zheng, O. Cetin, M. Y. Hwang, X. Lei, A. Stolcke and N. Morgan (2007). "Combining discriminative feature, transform, and model training for large vocabulary speech recognition", in *Proc. IEEE ICASSP2007*, Vol. 4, pp. 633-636, Honolulu, Hawaii, USA.

[13] T. Ng, B. Zhang, S. Matsoukas and L. Nguyen (2011). "Region dependent transform on MLP features for speech recognition", in *Proc. ISCA INTERSPEECH2011*, pp. 221-224, Florence, Italy.

[14] N. Cheng, X. Liu and L. Wang (2011). "Generalized variable parameter HMMs for noise robust speech recognition", in *Proc. ISCA INTERSPEECH2011*, pp. 482-484, Florence, Italy.

[15] N. Cheng, X. Liu and L. Wang (2011). "A flexible framework for HMM based noise robust speech recognition using generalized parametric space polynomial regression", *Science China, Information Sciences*, 54(2), pp. 2481-2491, 2011.

[16] Y. Li, X. Liu and L. Wang (2012). "Structured modeling based on generalized variable parameter HMMs and speaker adaptation", in *Proc. IEEE ISCSLP2012*, pp. 136-140, Hong Kong, China.

[17] Y. Li, X. Liu and L. Wang (2013). "Feature space generalized variable parameter HMMs for noise robust recognition", in *Proc. ISCA INTERSPEECH2013*, pp. 2968-2972, Lyon, France.

[18] R. Su, X. Liu, L. Wang (2013). "Automatic model complexity control for generalized variable parameter HMMs", in *Proc. IEEE ASRU2013*, pp. 150-155, Olomouc, Czech Republic.

[19] A. Bjorck and V. Pereyra (1970). "Solution of Vandermonde Systems of Equations", Mathematics of Computation (American Mathematical Society) 24(112): pp. 893-903.

[20] K. Fujinaga, M. Nakai, H. Shimodaira and S. Sagayama (2001). "Multiple-Regression Hidden Markov Model", in *Proc. IEEE ICASSP2001*, Vol. 1, pp. 513-516, Salt Lake City, Utah, USA.

[21] X. Cui and Y. Gong (2007). "A study of variable-parameter Gaussian mixture hidden Markov modeling for noisy speech recognition", *IEEE Transactions on Audio, Speech and Language Processing*, 15(4):1366-1376, 2007.

[22] D. Yu, L. Deng, Y. Gong and A. Acero (2009). "A novel framework and training algorithm for variable-parameter hidden Markov models", *IEEE Transactions on Audio, Speech and Language Processing*, Vol 17(7), pp. 1348-1360, 2009.

[23] Z. Ling, K. Richmond and J. Yamagishi (2013). "Articulatory control of HMM-based parametric speech synthesis using feature space switched multiple regression", *IEEE Transactions on Audio Speech and Language Processing*, vol.21, no.1, pp. 207-219, Jan. 2013. doi: 10.1109/TASL.2012.2215600.

[24] A. P. Dempster, N. M. Laird and D. B. Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society*, 39(1):1-39,1977.

[25] C. Runge (1901). "Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten", *Zeitschrift für Mathematik und Physik*, 46:224-243.

[26] G. Schwartz (1978). "Estimating the Dimension of a Model", *The Annals of Statistics*, pp. 461-464, Vol. 6, No. 2, February 1978.

[27] X. Liu and M. J. F. Gales (2003). "Automatic model complexity control using marginalized discriminative growth functions", in *Proc. IEEE ASRU2003*, pp. 37-42, St. Thomas, U.S. Virgin Islands.

[28] X. Liu and M. J. F. Gales (2004). "Model complexity control and compression using discriminative growth functions", in *Proc. IEEE ICASSP2004*, Vol. 1, pp. 797-800, Montreal, Quebec, Canada.

[29] X. Liu and M. J. F. Gales (2007). "Automatic model complexity control using marginalized discriminative growth functions", in *Proc. IEEE Transactions on Audio, Speech and Language Processing*, vol.15, no.4, pp.1414-1424, May 2007.

[30] W. Chou and W. Reichl (1999). "Decision tree state tying based on penalized Bayesian information criterion", in *Proc. IEEE ICASSP1999*, Vol. 1, 345-348, Phoenix, Arizona, USA.

[31] The Kaldi speech recognition toolkit. *http://kaldi.sourceforge.net*

[32] S. Young et al., "The HTK Book Version 3.4.1 ", 2009.

[33] R. Lippmann, E. Martin and D. Paul (1987). "Multi-style training for robust isolated-word speech recognition", in *Proc. IEEE ICASSP1987*, pp. 705-708, Dallas, Texas, USA.